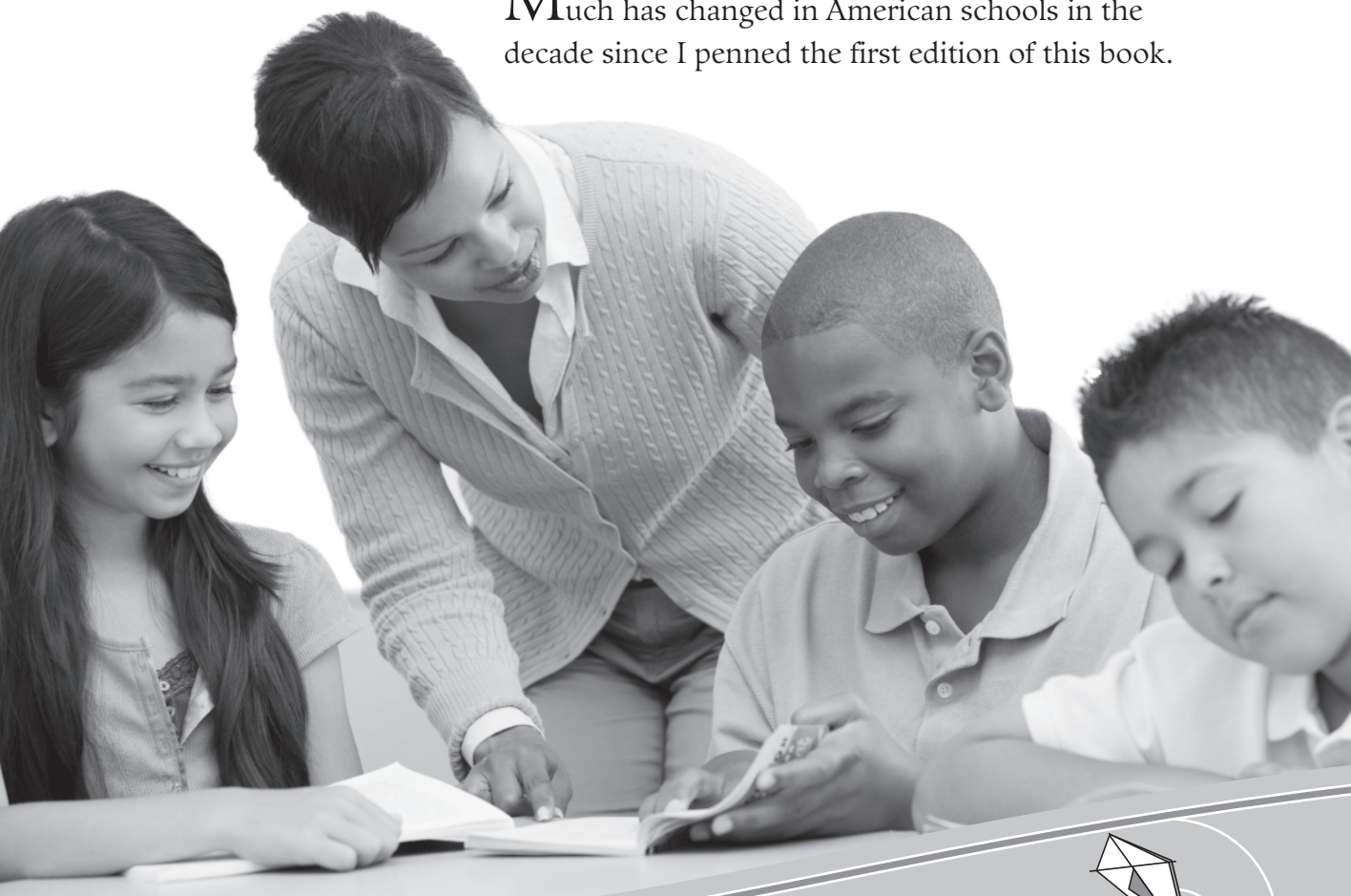
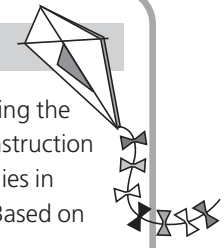
Chapter 1
Reading Achievement and Instruction in U.S. Schools

Much has changed in American schools in the decade since I penned the first edition of this book.



Most of that change was stimulated by the federal No Child Left Behind Act (NCLB) of 2001. We now know that the hundreds of millions of federal dollars spent under NCLB had one positive outcome: First-grade students from low-income homes enrolled in Reading First schools read nonsense syllables faster and more accurately than low-income first-graders in schools not benefiting from Reading First funding. In addition, children in Reading First schools received more minutes of reading instruction every day than the other poor kids and more of that reading instruction was focused on the five pillars of early reading instruction outlined in the report of the National Reading Panel. But, not surprisingly (from my point of view), at the end of first, second, and third grade there was no difference in

National Reading Panel



The National Reading Panel (NRP) was charged by Congress with recommending the scientific studies that were worthy of consideration in the design of reading instruction in the future. The NRP elected to examine only the experimental research studies in developing their report, a decision decried by many educational researchers. Based on their review of this body of research they concluded the following:

- Developing phonemic awareness and phonics skills in kindergarten and first grade was supported by the research but systematic phonics was not effective for struggling readers in grades 2 to 6.
- Providing regular guided oral reading with a focus on fluency was important.
- Silent reading was recommended for developing fluency, vocabulary, and comprehension skills (though the panel felt that the research reviewed had not adequately demonstrated the benefits of various incentive programs for increasing reading volume).
- Direct teaching of comprehension strategies was recommended and it was noted that providing good comprehension strategy instruction is a complex instructional activity. Thus, the panel recommended extensive, formal preparation in comprehension strategies teaching for all teachers.
- Little research was available to support the use of technology (e.g., computers) in teaching reading, but the few studies available suggested that it was possible that there was a potential for some benefits to students.

For further information, see Allington, R. L. (2002). *Big Brother and the national reading curriculum.* Portsmouth, NH: Heinemann.

the reading achievement posted by the children in these two types of schools, nor was there any difference in student engagement in reading (Gamse et al., 2009). Additionally, the reading gap between children from more and less economically advantaged families did not close in the past decade, even though that was the impetus behind NCLB. One could say NCLB was a failure in fostering better reading achievement.

So why did all this extra federal money and local reading instructional time not produce either better readers or narrow the rich/poor reading gap? As I have argued earlier (Allington, 2002a), I think it was because the NCLB wasn't designed to raise reading scores. Instead, it was simply part of a larger scheme intended to reduce the power of teacher unions, colleges of education, and teacher professionalism—all ultimately, I believe, in the name of greater privatization of public education in the future. If public schools cannot raise the reading achievement of poor children even when given substantial extra funding, then something else must be needed was the mindset of those who developed NCLB.

But with its focus on teacher and student accountability and penalties for schools failing to raise reading achievement and close the reading achievement gap, the No Child Left Behind Act was designed largely by politicians and policy makers, most of whom had never spent a day in a classroom as a teacher. The best you can say is they believed the hypothesis that greater accountability and sanctions for schools and educators would somehow improve reading achievement. They believed that schools were like widget factories, where such plans had improved productivity (and raised the profit margin). But as Diane Ravitch (2010) has so articulately described, there was no evidence that any of the components of NCLB had ever raised reading achievement, anywhere.

We are now entering roughly the fiftieth year of reforming American schools and attempting to close the reading achievement gap between children from economically different families while also raising the achievement of all children substantially. The NCLB Act is simply the legislative extension of the Elementary and Secondary Education Act of 1966 that was a major component of the War on Poverty declared by then-President Lyndon B. Johnson. But we have more poor families and children today than we had way back then. And we still have a large gap in the reading abilities of children from low-income and middle-class families. However, what we have today that was missing back then is reasonably clear evidence that we can teach virtually every child to read and have virtually all of them reading on grade level by the end of first grade (Mathes et al., 2005; Scanlon et al., 2005; Vellutino et al., 1996). The only children who fail to meet the grade-level criteria are those who fail to attend school regularly and those with the most severe disabilities.

In this book, then, I hope to convey to you, first, the promise of the power of effective reading lessons and, second, a solid description of just what you really have to pay attention to when designing reading lessons for struggling readers.

Why Another Book?

The rationale for writing this book is that much of the rhetoric and policy making that surrounds current efforts at "reforming" U.S. reading instruction are misguided. They are misguided because a considerable amount of the reform sentiment focuses on features of instruction that don't really matter that much in the grand scheme of things. Many of the reforms are narrowly conceived and simply cannot have the sort of impact that we might hope for given the time, money, and energy that has been spent. In this book I will refocus attention on the few things that really matter in teaching children to read (and the things over which we as teachers and administrators can actually exert a degree of control).

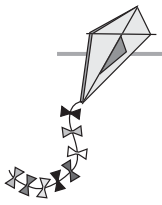
Why Now?

Simply, the reason for revising the book now is that we can now see the impact of the No Child Left Behind Act on the design of school reading programs. And I, for one, don't much like what I am seeing. I am writing because I am worried that, to date, "What the research says . . ." has been narrowly interpreted and focused almost wholly on the very beginning stages of reading instruction. I am writing because I am deeply worried that so much of what we have learned about teaching reading effectively—especially to children who have difficulty—is being routinely ignored. I am writing because the research is being misrepresented (see Allington, 2002b; Allington & Woodside-Jiron, 1998, 1999; Coles, 2003; Garan, 2002; Pearson, 2004; Taylor, 1998). I am writing because much of what might prove useful instructionally in first grade is being misapplied to older children and to children having difficulty. I am also writing because the scientific evidence has shown that those Reading First initiatives did not improve reading achievement even though teachers in the Reading First schools allocated more time for teaching reading than teachers in other schools.

My goal is to provide a readable, practical treatise on designing a more effective reading instruction. My long-standing concern for children who have difficulties learning to read will be evident because it is the instruction of those children that seems most often to go awry in schools.

But to begin, let me correct some of the misunderstandings about U.S. children's reading proficiency and U.S. reading instruction today.

So How Bad Is the Situation in Terms of Reading Achievement?



Actually, the answer to how bad the reading achievement situation is depends on your reference point. For instance, in a recent international comparison of children's reading achievement (Bracey, 2004), U.S. fourth-graders were ranked ninth in the world. Only three nations had scores that were significantly higher (see Table 1.1). U.S. ninth-graders ranked right in the middle, at the international average. These data often surprise many people in the United States, including educators, but as they say, "You can look it up!"

On these international assessments there were separate sections on prose and informational reading. Students from the United States performed sharply better on prose reading than they did on informational text reading. They ranked third on prose reading, with only a single nation earning statistically higher scores, and twelfth on reading informational texts, with five nations statistically ahead. But the U.S. best readers performed well, with twice as many students ranking in the top 10 percent of readers (19 percent versus 10 percent), as would be expected. Students in schools enrolling fewer than 25 percent free-lunch students performed well above the average score of the top nation. But schools with more than 75 percent free-lunch students ranked twenty-eighth with the score of 485 (Bracey, 2004).

Obviously, students in many U.S. schools read quite well, whereas students in other schools, primarily schools enrolling many children from low-income families, lag far behind. On average, though, the performance of both our 9- and 15-year-olds equaled or exceeded that of students in the majority of other industrialized nations participating in these international assessments.

National Assessment of Educational Progress

Every two years, the U.S. Department of Education releases a new Report Card on Reading. This report card draws national headlines along with, typically, statements of concerns from federal and state policy makers—concerns that U.S. schools are failing their responsibility to produce a literate citizenry. The report card details the findings of the National Assessment of Educational Progress (NAEP), a series of assessments covering an array of subject areas. The NAEP reading assessments are the most frequently administered and seem to generate the most discussion. The general theme of such discussions recently has gone something like this: "We've dramatically increased education expenditures, but reading scores remain flat with huge achievement gaps between different subgroups." In this section, I will attempt to help you understand that the general interpretation of NAEP achievement data ignores important gains that have been made over the past 40 years.

TABLE 1.1 Combined Reading Literacy Scores, Ages 9 and 15

Age 9		Age 15	
Sweden	561	Finland	546
Netherlands	554	Canada	534
England	553	New Zealand	529
Bulgaria	550	Australia	528
Latvia	545	Ireland	527
Canada	544	Korea	525
Lithuania	543	England	523
Hungary	543	Japan	522
United States	**542**	Sweden	516
Italy	541	Austria	507
Germany	539	Belgium	507
Czech Republic	537	Iceland	507
New Zealand	529	Norway	505
Scotland	528	France	505
Singapore	528	**United States**	**504**
Russian Fed.	528	Denmark	497
Hong Kong	528	Switzerland	494
France	525	Spain	493
Greece	524	Czech Republic	492
Slovak Rep.	518	Italy	487
Iceland	512	Germany	484
Romania	512	Lichtenstein	483
Israel	509	Hungary	480
Slovenia	502	Poland	479
Norway	499	Greece	474
Cyprus	494	Portugal	470
Moldova	492	Russian Fed.	462
Turkey	449	Latvia	458
Macedonia	442	Luxembourg	441
Colombia	422	Mexico	422
Argentina	420	Brazil	396
Iran	419		
Kuwait	396		
Morocco	350		
Belize	327		

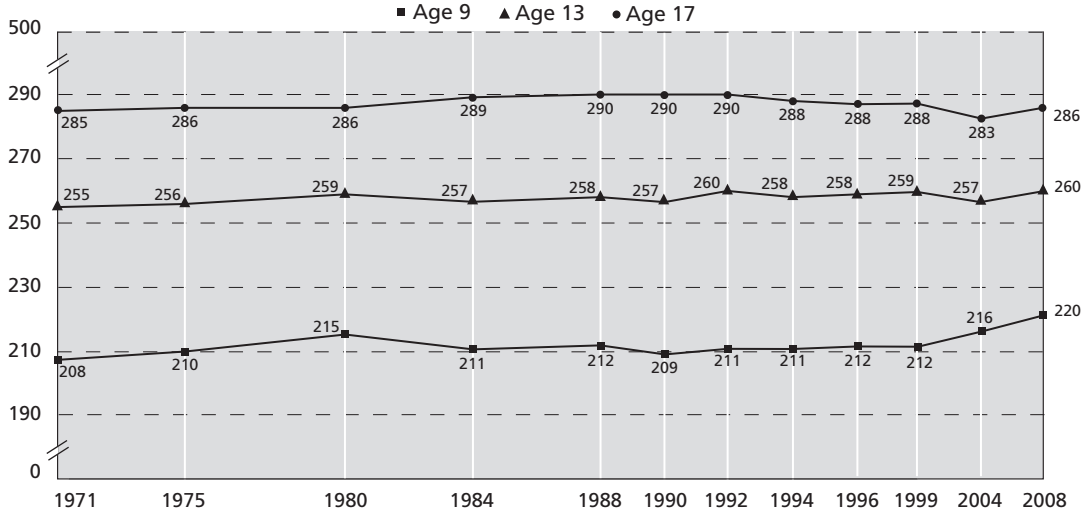
Source: National Center for Education Statistics. (2004). *The nation's report card: Reading highlights 2003*. Washington, DC: U.S. Department of Education, Institute for Education Sciences.

(For full information on the NAEP, including sample test items and state NAEP reports, visit http://nces.ed.gov/nationsreportcard.)

There are actually two NAEP assessments. One, the Trend Assessment, is designed to allow comparisons of reading achievement over time. The most recent Trend Assessment was completed in 2008. Figure 1.1 shows the longer-term pattern of reading achievement for each of the three grade levels tested. A quick glance suggests that scores have remained largely stable for 40 years. There is a worrisome small decline in twelfth-graders' reading performance since 1990 and similarly a small rise in fourth-graders' reading performances in that same period.

Bracey (2004) notes, however, the presence of Simpson's Paradox in the NAEP Trend Assessment data. Simpson's Paradox illustrates how average achievement data reports can obscure important findings. For instance, since 1971 the average NAEP trend reading score for fourth-grade Black students rose 34 points (from 170 to 204), Hispanic students' scores rose 24 points (from 183 to 207), and White students' scores rose 14 points (from 214 to 228). Similar gains were made at eighth and twelfth grades. In other words, across the 40-year period, much progress was made in closing the achievement gap between White and minority students. But the overall average gain in the NAEP fourth-grade trend data shows an average reading improvement of only 12 points between 1971 and 2009 (from 208 to 220) and only a 5-point gain since 1980 (from 215 to 220). How can it be that every subgroup grew by a greater amount than the average gain reported?

FIGURE 1.1 NAEP Reading Long-Term Trend Scores for the Nation

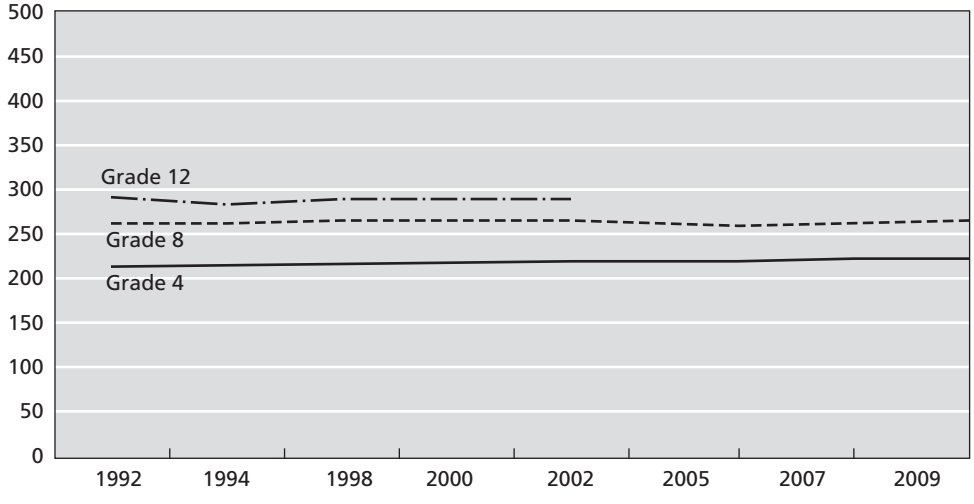


Source: National Center for Education Statistics. *National Assessment of Educational Progress.* Retrieved from http://nces.ed.gov.

This is where Simpson's Paradox comes in. The average trend score comparisons are a bit of an apples and oranges comparison. What I mean is that minority enrollment expanded rapidly in U.S. schools over the 1971 to 2008 time period, from about 15 percent to about half (44 percent) of the student population tested. Even with their improved reading performances, minority students still trail White students by a substantial margin (a gap of 20 to 30 points). Thus, the population shifts resulted in about three times as many minority students in the current NAEP assessment pool as there were in the earliest NAEP assessment pools. The overall lower achievement of this much larger group of minority students produced a trend line that looks almost flat over the 40-year period. But underneath the overall average trend line are substantial improvements by every subpopulation of students, especially minority students. A worrisome trend, however, is that although the racial and family income achievement gaps narrowed over time, those gaps are still far too large, and it was these gaps that fueled the No Child Left Behind legislation. But NCLB had no effect on closing this reading achievement gap.

The second NAEP assessment offers the Main Assessments Report (National Center for Educational Statistics, 2009). These are the ones that are commonly discussed. Because of changes in the NAEP assessment and in the administration (including accommodations for pupils with disabilities, for instance), Main Assessment Reports are more difficult to compare over time. The National Assessment Governing Board (NAGB) indicates that only NAEP Main Assessments from 1992 to 2008 can be considered equivalent (see Figure 1.2).

FIGURE 1.2 NAEP Main Report Reading Scores 1992–2009



Source: National Center for Education Statistics. (2008). *The nation's report card: Reading highlights 2008.* Washington, DC: U.S. Department of Education, Institute for Education Sciences.

The NAEP scores over the past two decades have shown little change, with only small improvements across the three grades (4, 8, and 12). The percentage of students failing to achieve the basic level of reading performance dropped from 40 to 37 percent at fourth grade and from 31 to 26 percent at eighth grade. The results also indicate that, in general, the best readers are reading better and the worst readers are reading about the same or slightly worse than was the case two decades ago. The gap between White and minority students and the gap between more and less economically advantaged students have not narrowed in the past two decades. In other words, the progress noted previously in closing achievement gaps occurred between 1971 and 1990. Since then, little progress has been made in overcoming this challenge, even though this was a major goal of federal legislation.

The achievement of U.S. elementary and middle school students on nationally normed, standardized commercial tests of reading achievement have been rising since 1980. Between 1965 and 1980, the heyday of the basic skills instruction movement, these scores declined but began to rise substantially in recent years (Bracey, 2004). For instance, on the Iowa Test of Basic Skills, the average fifth-grader's achievement in 1990 roughly equaled the average sixth-grader's achievement in 1975, and the average third-grader's achievement in 1990 equaled the achievement of the average fourth-grader in 1955. Across the 40- to 50-year period (1960–2010), elementary school student achievement rose quite dramatically, whereas average middle school achievement improved only modestly. But at all grade levels children today outperform children from earlier eras of U.S. schooling.

Finally, one other indicator that might be used is the readability formulas that were created to estimate the difficulty of books. The oldest and most popular of these formulas originated in the 1940s and 1950s. However, two such stalwarts—the Dale-Chall and the Spache readability formulas—were renormed in the 1980s because they no longer accurately reflected the grade difficulty of texts. In both cases, the difficulty estimates were overestimating the complexity. So, in renorming, what had been a seventh-grade-level book became a sixth-grade-level book.

Various explanations for the wave of negative information that has filled the media have been offered, but a simple principle for educational reporting—good news is no news—may provide the simplest explanation. Of course, when advocates of privatization of education attempt to move their political agendas, bad news about U.S. schools is necessary as a lever to attempt to persuade the public to accept such a radical shift in the financing of public education (Bracey, 2004). And for three decades (since 1981), the White House has been occupied by privatization advocates. But, interestingly, the U.S. public seemed more confused than convinced by the notion that U.S. education was failing.

In one Gallup poll (Coles, 1999), half of U.S. adults awarded their local schools an A or B grade for effectiveness. However, fewer than half of these same adults awarded U.S. schools generally the same grade. In other words, lots of folks believe their children attend one of the few good schools in the country. These adults

award grades to local schools based on more direct knowledge of the achievements of those schools. They award lower grades when asked to rate schools that they have little direct knowledge of. These are schools that they acquire information about only indirectly, largely through media accounts. So it seems that most U.S. adults believe that *other* schools need to improve but *their* schools are doing a creditable job of educating children.

But if Achievement Is Actually Rising, Why Another Book?

Research points to high reading achievement levels in U.S. students today—no doubt about it. But research also points to several disturbing trends in U.S. reading achievement. The first is the trend for certain groups of children to lag behind their peers in literacy learning. The more disturbing part is that these children are all too often predictable. For instance, researchers at the Rand Institute on Education and Training (Grissmer et al., 1994) found that students whose parents were not high school graduates had achievement levels significantly below the achievement levels of children whose parents were college graduates. Family income was also related to achievement, as were mother's age (with the children of older mothers achieving higher levels) and other factors. But, perhaps surprisingly, these researchers found little relationship between achievement and children from single-parent homes or homes where both parents worked. They concluded that only parent educational levels and family income were related to achievement.

There are also other group differences in achievement. The one most frequently mentioned is the Black/White achievement gap. The 2009 NAEP results (National Center for Education Statistics, 2009) show a 25-point gap between Black and White fourth-graders' reading scores and a 26-point gap at eighth grade. These are sizable differences. For instance, the gap between Black eighth-graders and White fourth-graders is 14 points (246 and 230).

But there are other gaps as well. The Hispanic/White gap is almost identical to the Black/White gap at both grade levels. There is a rich/poor gap of 28 and 26 points at the fourth- and eighth-grade levels, respectively, again similar to the Black/White gap.

Then there is the boy/girl gap. This gap is smaller, but girls outperform boys in reading by 7 and 10 points at grades 4 and 8, respectively. This gender gap was also observed in the international achievement comparisons, with girls significantly outperforming boys in reading in every nation.

There is also one additional confounding factor: Retention in grade. Flunking children with low reading scores has been mandated in several states and in a

number of city school districts. This means that we can now find many children who should be in middle school still enrolled in the elementary grades. In one of our most current studies (Allington & McGill-Franzen et al., 2010) we found 13- and 14-year-old students enrolled in grades 4 and 5. Recently we had a discussion with a 17-year-old seventh-grade student! By state mandate this student had spent 3 years in third grade and 3 years in sixth grade along with having repeated two other grades. Even if this student is promoted every year onward, he will be 22 years old when he earns his high school diploma—assuming of course that he stays in school, which seems unlikely.

I mention these cases because the children taking the fourth- and eighth-grade NAEP reading assessments are getting older than they used to be. If a child has repeated grade 3 three times and is now 12 or 13 years old, shouldn't her reading scores be higher than had she been promoted and tested when she should have been in fourth grade? There are simply too many students taking the fourth-grade NAEP today who should be in grades 5, 6, or 7 for anyone to claim that fourth-grade reading scores are improving.

One way to consider these data is as additive risk factors. Being poor places you at risk of reading difficulties; being a boy or being a minority also places you at risk. One might examine local data using these broad categories. Do more boys, more poor boys, more poor minority boys struggle with reading? Is this also true for flunking a grade? If so, this seems to say more about problems in the school system than it says about these boys.

Three Challenges

The evidence indicates that U.S. schools currently work better for certain children than for others. In order to hope to fulfill the promise of public education, schools must work for all children—regardless of gender, race, or which parents the children got. One challenge for U.S. education is designing schools that are less parent dependent, where all children can expect to be successful readers and writers. Given that only about one-third of all entering kindergarten students arrive at school not knowing all the letter names of the alphabet (Zill & West, 2001) and that about one-third of all fourth-graders fail to achieve the basic level of reading proficiency, it becomes clearer that schools don't work well for some children. These are the children of low-income families where levels of parental educational attainment are also typically low. If our schools fail to teach these children to read well, how likely is it they will ever become productive citizens?

There is a second challenge for U.S. education. Although we have been largely successful in teaching children to read and write at basic levels of proficiency, the vast amount of growing technology places higher-order literacy demands on all of us. As Bill Kovach and Tom Rosenstiel point out in their hard-hitting book, *Warp*

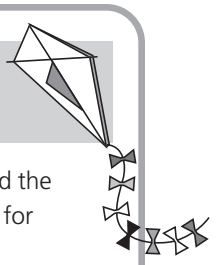
Speed: America in the Age of Mixed Media (1999), we have entered an age of unfettered information flow. Historically, only a few large publishers could afford to provide national news and information dissemination. These authors argue that with that power came a sense of responsibility for attempting to ensure an accuracy and completeness of the information—in other words, journalistic codes of ethics and editorial decisions about the quality of the "evidence" supporting a story. They note that the modern information environment is different with its proliferation of news outlets, 24-hour news, and infotainment channels. They point to a new "journalism of assertion" as the dominant mode of delivery. This mode has fewer checks and balances and literally places far greater demands on the reader, viewer, and listener. These demands include synthesizing and evaluating information from multiple sources. These multiple information sources have fewer editorial controls and fewer filters through which information is sifted for accuracy, reliability, and civility. And Kovach and Rosenstiel hardly even mention the Internet but instead focus on the television and print media.

The Internet imposes virtually no controls on information quality and reliability. Type the word *Holocaust* into an Internet search engine and you will find many web pages denying the Holocaust ever took place. Although materials denying the Holocaust have been around since the 1950s, never have they been so widely accessible to so many people; never have they appeared so "official."

Because of the increase in the unfettered flow of information, U.S. schools need to enhance the ability of children to search and sort through information, to synthesize and analyze information, and to summarize and evaluate the information they encounter. On the one hand, the performances of U.S. students on the NAEP have been improving, having risen to historically high levels of attainment. On the other hand, only a few U.S. students seem to be able to demonstrate even minimal proficiency with higher-order literacy strategies (and the children most likely to demonstrate these are those children whose parents have high levels of educational attainment). Even a quick examination of the NAEP items that so many children and adolescents find especially difficult suggests that we have done a better job of teaching the basic literacy skills (word recognition, literal comprehension) than the higher-order thinking skills and strategies. The items attempting to assess higher-order proficiencies do not require rocket-scientist–type performances. Many, in fact, require little more than the types of judgments about information, ideas, and assertions that an adult makes nearly every day.

There is a third challenge for U.S. education that needs to be mentioned. Our schools create more students who *can* read than students who *do* read. Too many students and adults read only when they are required to. Interest in voluntary reading begins to fall in the upper-elementary grades, declines steeply in middle school, and continues to fall across high school. We seem to be producing readers who can read more difficult texts but who elect not to read even easy texts on their own time.

Sample Items from a Recent Fourth-Grade NAEP



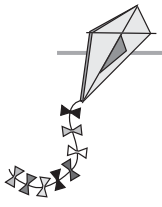
After the students read a 2½-page Ashanti folktale entitled "Hungry Spider and the Turtle," the following written response questions were posed (National Center for Education Statistics, 2001):

- There is a saying, "Don't get mad, get even." How does this apply to the story?
- Who do you think would make a better friend, Spider or Turtle? Explain why.
- Think about Spider and Turtle in the story. Pick someone you know, have read about, or have seen in the movies or on television and explain how that person is like either Spider or Turtle.

In each of these short written responses students had to *think* about the story, not just recall the story (though recalling characters and their traits noted in the story is obviously essential to responding adequately). The point to be made is that being able to locate or remember the correct answer (word or phrase) to a multiple-choice item, the traditional measure of comprehension, is simply much less demanding than responding to these new measures of reading comprehension.

U.S. schools, especially elementary schools, produce children who rank among the world's best readers. The schools are improving. More children are better readers than ever before, but there are still substantial challenges that need to be confronted.

Can Educational Research Provide Support for Meeting These Challenges?



The answer to this question is *perhaps.* We have learned much about what sorts of schools, classrooms, and lessons foster reading proficiency. We have learned about how to redesign schools and classrooms and lessons to better meet the needs of children who find learning difficult (Allington & Cunningham, 2006). We have also been learning more about what sorts of classrooms and lessons foster the higher-order literacy that seems often neglected (Allington & Johnston, 2002).

And we have learned more than a little about the sorts of classrooms and lessons that foster ownership of reading and writing (Guthrie & Humenick, 2004). But with all we have learned, there still is no simple blueprint for restructuring schools, classrooms, and lessons. Perhaps that is because blueprints differ for virtually every building constructed. Perhaps the same is true of school restructuring. Because we build different buildings for different clients, we may need to build different classrooms and lessons for different students in order to meet the challenges that confront us.

In suggesting that educational research can provide no blueprint I am not suggesting that educational research can provide no guidelines—quite the contrary. But educational research is a slippery beast. There is a trove of published studies and an even greater supply of unpublished studies to be found in ERIC, in Dissertation Abstracts, and on publishers' websites. Much of the educational research available, even in the published papers, fails to meet rigorous quality criteria. Skeptics suggest that "You can prove anything with research." To some extent that is true, especially at the level of comparing particular programs, materials, or methods. Such comparisons have a long history in education even though the results are rarely consistent from study to study. Too often, proponents of a particular method, material, or program selectively review the available research and report on studies supporting their biases. A more comprehensive review often shifts the resulting conclusions.

It is not uncommon to hear someone initiate a discussion of any particular method, material, or program with the phrase "Hundreds of studies show . . . ." Truth be told, it is impossible to locate 100 studies showing the same effect for any method, material, or program. Consider that Jeanne Chall, author of *Learning to Read: The Great Debate* (1987), located fewer than 100 studies comparing different approaches to teaching beginning reading even though her review covered 80 years of research. Every approach she examined produced the best results in at least one of the comparison studies, and every approach was found less effective in some studies. Similar findings after 27 coordinated studies compared methods and materials in the largest beginning reading field experiment ever conducted led the authors of the *First Grade Studies* (Bond & Dykstra, 1967) to conclude: "Children learn to read by a variety of materials and methods. . . . No one approach is so distinctly better in all situations and respects than the others that it should be considered the one best method" (p. 75).

One useful component of NCLB was the creation of the federal What Works Clearinghouse (WWC). This agency was tasked with reviewing educational research and developing a sort of score card in different areas. They reviewed the research available on over 150 reading programs. But only one reading program received the WWC's rating as having "strong evidence" of improving reading achievement. That program was Reading Recovery, an individualized reading tutorial for first-grade students having difficulty. Only two other commercially

available reading programs were rated as having "possible evidence" of positive effects. Not a single core reading program had even one study supporting its use. And almost no reading program currently available had any evidence it improved reading achievement. Some had evidence they helped children read faster or decode nonsense syllables better but only one had strong evidence it improved reading achievement. Nonetheless, commercial vendors routinely tout their programs as "research-based" and typically have glossy slides showing how well their program works. Visit the What Works Clearinghouse website (www.whatworks.ed.gov) for full details on the findings as well as plain language material on making sense from research claims.

On most educational questions there are only a handful of published studies, and often these are smallish and local. In the case of materials and programs, the number of published studies shrinks even further. Additionally, more often than not, the majority of the few published studies are authored by the developers and marketers of the materials and programs. In other words, there are few independent evaluations of most materials and programs available. Several examples come to mind.

First, there was the Waterford Early Reading program, an expensive, computer-based package designed for kindergarten and primary-grade classrooms. Although the publisher provided data from a number of local unpublished evaluation studies, the only independent experimental field trial found no effect on early reading achievement. The authors (Patterson et al., 2003) concluded:

> *It is clear from the present results that the Waterford Early Literacy Program had relatively little overall effect on the participants' literacy development. . . . [T]o the extent there was an effect, it was in a negative direction. . . . On the other hand, teacher variables had a consistently strong effect on reading success.* (p. 198)

Then there is the heavily promoted Success for All (SFA) program and the highly publicized Direct Instruction (DI) materials. In both these cases there exists a substantial set of studies, often published in professional journals (with fewer studies published in peer-reviewed research journals). In both programs there are some independent research studies that have also been published. The sets of SFA studies generally find that the program produces statistically significant achievement effects when SFA schools are compared to control schools (schools without SFA). The independent studies report the same sort of effects, though often reporting smaller differences in achievement between schools than the studies by the SFA developers. There exists, then, published evidence that implementing the SFA program improves achievement.

However, as Venezky (1998) points out in his reanalysis of the SFA data, the program continues to produce large numbers of children with dismayingly low reading achievement. He reported that fifth-grade students in the SFA schools

had reading achievement levels at the middle–third-grade level compared to the beginning–third-grade reading levels of fifth-graders in the control schools. Venezky does not question whether SFA produced higher achievement levels—it did. Rather, he asks whether we are willing to accept *so little* improvement in reading as sufficient evidence that the SFA program should be recommended for wide implementation.

In the case of the DI materials, the body of research extends back 40 years, and the evidence has been controversial across that period of time (e.g., House et al., 1978; Jordan, 2005; Schweinhart & Weikart, 1998). The majority of the DI research has been done by the developers, with much of it reported in *Effective School Practices,* the in-house magazine of the Association for Direct Instruction edited by one of the DI program authors. This research was summarized in a self-published selective review (Adams & Englemann, 1996)—a review that omitted a number of DI studies that did not report positive effects. There are so few independent studies published in peer-reviewed research journals that some scholars largely discount the evidence available (Stahl et al., 1998).

This sort of criticism could just as easily be offered on Wilson Phonics, whole language, and the Houghton Mifflin (or any other major publisher) reading program. The important point here is that "What the research says . . ." is currently an almost meaningless phrase. In other words, virtually every proponent of any method, material, or program can find some sort of evidence to prove what they have to offer works somewhere, some of the time. By selectively reviewing the evidence, by creating magazines to publish their own supportive data (because no peer-reviewed journal would accept it as unbiased), and by controlling the design of the evaluation and the implementation of their favorite method, material, or program, almost anyone can create the impression that "research shows" positive effects for their product or pedagogy.

Ideally, research studies would demonstrate the longer-term impact of interventions as well as report the shorter-term effects. Unfortunately, rather few studies report effects over periods greater than one year, and many report the effects after only a few months. Longer-term studies are more complicated and more expensive but they are also essential.

There is a long-standing federal enthusiasm for packaged reading reform. Unfortunately, we also have 50 years of research showing that packaged reforms simply do not seem reliable in improving student achievement (Allington & Nowak, 2004). Nonetheless, federal and state education policy makers ignore these studies and attempt to reform from the top down. Sometimes the push is for a particular method of teaching children to read. Other times the push is for particular commercial reading programs. Much of the recent focus on so-called proven programs can be traced back to the early 1990s, when Congress passed the Comprehensive School Reform Development Act of 1990 (Berends et al., 2002). This legislation provided funds for adopting one of several national school reform models (e.g.,

Success for All, Accelerated Schools, America's Choice). These were supposed to be "proven programs" that supported schoolwide reform. The RAND research group produced a 10-year evaluation of this federal effort. It concluded, "The initial hypothesis, that by adopting a whole-school design a school could improve its performance was largely unproved" (Berends et al., 2002, p. 173). In other words, while there were some schools that adopted one of the national designs and then saw student achievement rise, there were at least as many schools that adopted one of the designs in which achievement did not improve.

Publishers and promoters of packages and products have simply modified their advertising and advocacy campaigns to include the message that "research shows" that their product works. The education marketplace is a "buyers beware" market. Little scientific research exists that demonstrates any package or program works consistently and reliably. For instance, none of the "proven programs" that generated so much excitement more than a decade ago has withstood the independent research review. None of the commercial reading series has either. There is a lot of money to be made in the education marketplace, but that means convincing administrators and teachers to buy your stuff (including this book and my consulting services).

There is one final problem with research available today. In most cases, the available studies evaluated the effects of a literacy intervention on assessments of basic literacy, not on thoughtful literacy assessments. In other words, most studies use word lists, tests of subskill knowledge, or assessments of low-level comprehension found on traditional standardized tests with their multiple-choice items. We have only a handful of studies that have evaluated interventions against student attainment of the new, higher-order literacy standards. But it is against student achievement of these new, higher standards that schools are typically now being evaluated.

For years, no one actually paid much attention to "research" evidence on various methods, materials, and programs. Yes, marketing departments often created some flyer or glossy brochure designed to convince the occasionally wary buyer that there was a research base for the product. (And none of this is to suggest that research did not influence the design of educational methods, materials, and products—it did.) But today there are greater demands on publishers and promoters to have actual research on the effects the implementation has had on the achievement of students. This is different from being able to point to studies that influenced the design of a method, material, or program—the more traditional test. So, why the demand for research evidence now?

The No Child Left Behind Act of 2001

Congress passed the No Child Left Behind Act (NCLB) in 2001 with full bipartisan support. This legislation might be best viewed as an intensification of federal education policy, particularly policy focused on instruction in high-poverty schools. The NCLB law was an extension and reauthorization of the Elementary and Secondary Education Act of 1966 (ESEA). The ESEA marked the first real involvement of the federal government in the instructional arena. As part of the Great Society programs instigated by then-President Lyndon Johnson, ESEA was controversial. One reason was that the U.S. Constitution gave states responsibility for education. The ESEA represented a major new federal initiative that was viewed by many as a threat to local educational control.

A key component of ESEA was Title I of that act. Title I focused on providing high-poverty schools with additional funds to support supplementary reading instruction. The more recent Reading First component of Title I was a further effort to address the needs of children enrolled in the highest poverty-stricken schools. In both Title I and Reading First schools this additional federal funding was specified as additional reading lessons, "in addition to" the reading instruction already available to children in the school. Title I funds were to be used to "supplement" the classroom reading lessons. Thus, Title I services were often designed as pullout reading lessons in which children left the classroom for an additional 30 minutes of small group reading instruction after completing their 90-minute classroom reading lessons. In many states this additional instruction was to be provided by certified reading specialists, teachers with additional graduate preparation in teaching reading to struggling readers.

However, in far too many states paraprofessionals or teachers with no graduate training in reading disabilities provided the extra lessons. In far too many schools paraprofessionals simply monitored struggling readers while they worked on some form of computer-assisted reading lessons. (This was done even though the research indicates that neither paraprofessional reading lessons nor computer-based reading lessons have accelerated reading development.) In other words, in far too many schools the design of these supplementary reading lessons contradicted what the research says about what sort of reading lessons might actually solve struggling readers' reading difficulties. That is why I noted on page 1 of this book that I was not surprised that the Reading First initiative failed to improve reading achievement (Allington, 2009b).

The design of additional reading instruction has not typically been research-based (Allington & McGill-Franzen, 1989; Vaughn & Linan-Thompson, 2003). Often, Title I and special education reading lessons replaced, in whole or in part, classroom reading lessons. The Title I legislation created the now-widespread "second system" of education (Allington, 1994b). This system is now represented

by all those efforts that are not viewed as a component of the general education classroom program (remedial reading, special education, bilingual education, migrant education, gifted education, and so on). Most of these second-system educational efforts were also created by federal legislation and each had its own regulations and separate funding stream. This design, I've argued, resulted in a quite fragmented effort to design and deliver more and better reading instruction to children who struggle with learning to read (Allington, 2006).

The current situation in most schools, districts, and states is one of continued fragmentation of instructional plans for providing extra reading support. This fragmentation is best observed, perhaps, in cases of children who qualify for multiple programs. For instance, a child who is from a low-income family, is not from a home where the primary language is English, and is identified as having a learning disability would qualify for extra instructional support from at least three federal programs in most schools. But there are few schools in which a child would actually receive the full set of services and almost no schools where those services would be well coordinated with each other and with the classroom instructional program.

No Child Left Behind has changed none of this. The Reading First component of NCLB is largely an extension of Title I remedial-reading services. But NCLB brought changes that expand federal influence on the design and delivery of remedial reading instruction. The Title I legislation had long required schools, districts, and states to submit comprehensive plans for how federal funds would be used to expand reading instructional services to eligible children and adolescents. It had long required a testing program to demonstrate that federal dollars were benefiting struggling readers in high-poverty schools. The NCLB law expanded the specificity of both testing efforts and the nature of the instruction offered (Allington, 2002b).

Reading instruction in NCLB schools, at least, was to be based on "scientifically based reliable, replicable research," or SBRR. Unfortunately, little research was actually consulted when designing NCLB or in designing the reading instruction under NCLB. Instead, great hopes were placed on schools selecting commercial reading materials, materials that had a scientific base. Thus, in most states there was a push for adopting a single commercial reading program and then using that program "with fidelity." Two problems existed that the NCLB designers ignored. First, there was no research supporting the use of any of the core reading programs and none that supported the use of any supplementary reading program, except Reading Recovery. This was the message that the What Works Clearinghouse delivered when it released its report on research on reading programs. However, by the time the report was released, at least five years of mandated core reading programs use had been implemented. Second, what the research seems to suggest is that adaptation of commercial reading programs, adaptations based on student needs, produces better results than just following the reading program guidelines. But such adaptations are made only by teachers who are reasonably expert about both reading instruction and the children they are teaching (Duffy, 2004).

Thus, for at least the past decade U.S. schools have been implementing reading plans that no research supports—and doing so under the guise of "scientifically based reliable, replicable research." Congress has terminated funding for the Reading First portion of Title I funding, at least in part because the large federal study of the effects of the Reading First program indicated that no reading achievement gains occurred in Reading First schools (Gamse et al., 2009). In addition to finding no positive effects on achievement, Congress also noted the education department report from the Inspector General's office that indicated widespread corruption in the Reading First offices at the highest levels (Office of the Inspector General, 2006, 2007). The "corruption" involved recommendations for programs and assessments that had no research base, but these programs and assessments did have authors who were involved in making the decisions about what programs and assessments should be used. *Buyer beware* should have been the motto of NCLB.

Adequate Yearly Progress

Even though the earlier federal Title I legislation required accountability testing to gain federal funds, the testing plans differed substantially from state to state and district to district within states. Nonetheless, the Title I testing requirement was seen as promoting a much wider use of standardized tests, especially in high-poverty schools (Timar & Kirp, 1987). The NCLB expanded testing and reduced the variation in grade levels tested by requiring annual assessments of all children in grades 3 through 8. Also, the NCLB mandated annual assessments to determine whether various subgroups were making adequate yearly progress (AYP) in reading. Each state submitted plans for how AYP was to be met, but at the root of all plans was comparing the test scores of groups of children to the state standard.

Although earlier Title I programs also required year-to-year test score comparisons, fewer students were required to participate in the testing and a demonstration of achievement growth roughly comparable to one year's gain in the tests was the general goal. Of course, if struggling readers began the year below grade level, gaining a year in growth meant they still remained behind. The NCLB focuses on closing the achievement gap between various subgroups, and thus the goal is to accelerate reading growth such that struggling readers grow at a rate greater than one year per year. Acceleration then means that struggling readers must demonstrate achievement growth greater than the growth expected of children who read on grade level.

The NCLB requires that each state develop a schedule that ensures that all children will be reading at a targeted level by 2014. In general, the targeted level of reading proficiency is linked to each state's standards. But state standards for proficient reading vary widely. Some states have set low standards and others have set substantially higher standards. Thus, the proportion of students failing to meet AYP goals varies dramatically from state to state. An unintended impact of the NCLB law

is that some states lowered their proficiency standards so that the AYP goals will be easier to meet. In addition, states have established different dates when proficiency levels must be met. So in some states, few schools fail to meet AYP standards, whereas in other states, 9 out of 10 schools already fail to achieve AYP (Ryan, 2004; Peterson & Lastra-Anadon, 2010).

Currently, the NCLB is under consideration for reauthorization by Congress. The Obama administration is working hard to get states to raise their standards in both reading and math, and so far have indicated no substantial changes in the law, including the requirement to meet adequate yearly progress goals by 2014. I have no idea when Congress will reauthorize NCLB, whether Congress will change the name of NCLB, or whether Congress will substantially alter any of the requirements of NCLB. It is clear, however, that Congress is aware that several large-scale research studies (c.f., Mathes et al., 2005, Phillips & Smith, 2010; Vellutino et al., 1996) have demonstrated that every child can be reading at grade level by the end of grade 1 and be kept on level through the end of grade 3. Thus, I expect Congress to produce a new version of NCLB that has even more rigorous standards and requirements. Time will tell.

Disaggregation of Proficiency by Subgroups

The NCLB legislation requires annual testing of reading achievement and requires that the achievement of subgroups of struggling readers be accelerated to close any existing achievement gaps between these subgroups. Thus, schools must show that the achievement gains of poor children, minority students, and pupils with disabilities are narrowing any gap between the reading proficiency of these subgroups and the larger school population.

Schools must demonstrate that with each ensuing year the achievement of all subgroups is increasing and that any gap in the achievement of the subgroups and the majority student population is narrowed and finally eliminated. To achieve this end, federal funds are used to accelerate the reading development of struggling readers. Thus, schools focus on designing interventions that accelerate the reading development of children from low-income families, for instance, such that their achievement soon equals that of middle-class students. Or schools may design an intervention to accelerate the reading development of minority students to ensure that their achievement is soon on par with nonminority students. The education of pupils with disabilities has often been the responsibility of special education programs, but NCLB now includes those children as one of the targeted subgroups. Thus, schools must now also focus on accelerating the reading development of those children as well.

This is, perhaps, one of the real shifts in the federal model. This marks the first foray by the federal education agency to track academic outcomes for pupils with

disabilities (Gartner & Lipsky, 1987). Furthermore, tracking outcomes is not all that is required: In addition to the NCLB mandate that pupils with disabilities participate in the testing, it is also required that special education services produce achievement outcomes comparable to, or exceeding, those of pupils enrolled in general education. This largely represents a new conception of the role of special education instructional programs. The services now provided to special education students must accelerate achievement or risk having the school fall into the program improvement category. Of course, failure of any of the subgroups will produce that same outcome, but many students in other subgroups were already part of the federal accountability pool.

As has already been discussed, achievement gaps currently exist between each of the subgroups and the majority, middle-class population. The NCLB legislation garnered broad bipartisan political support because of its focus on eliminating such achievement gaps. I think the NCLB requirement to disaggregate achievement data by subgroups is a good idea. The 10-plus–year timeline for achieving comparable outcomes was a reasonable time period for implementing a revised design for remedial interventions that would be much more powerful than those now available in many schools. However, there are numerous problems with other aspects of No Child Left Behind—problems that may be legislated away over time—but problems nevertheless. One set of problems derives from the NCLB school improvement plans.

School Improvement under NCLB

Section 1116 of the NCLB law sets forth a number of complicated mandates that apply to schools where adequate yearly progress goals for subgroups are not being met. These mandates will be presented as they appeared in the original legislation. As already noted, these plans and mandates may be legislatively modified over time, if only because the impact will likely be widespread and in many cases quite severe. A number of state legislatures and various governors have called for substantial changes in the school improvement mandates.

The key to avoiding falling under the school improvement guidelines is to have all subgroups meet reading achievement goals; schools need to demonstrate that the federal funds they have received were allocated to interventions that accelerated reading achievement of poor, minority, English language learners, and pupils with disabilities such that the existing achievement gap is narrowed and, ultimately, eliminated. Just how much reading growth is necessary depends on how far behind struggling readers in the subgroups might be and the nature of the schedule each state has set for eliminating the gap. But it seems likely that reading growth of more than a year's gain will be needed in each subgroup to avoid the corrective actions the law mandates.

Under NCLB, each school district is required annually to review the progress of every school receiving funding from NCLB and to report to the public the results of the testing and whether the school met AYP goals. When any school fails to achieve AYP for two consecutive years, that school must develop a school improvement plan that covers a two-year period, incorporate scientifically based research to address specific problem areas that caused the failure, and adopt policies and practices for core academic subjects that have the greatest likelihood of raising student achievement to state proficiency levels.

In addition, the school improvement plan requires that not less than 10 percent of its federal funds be allocated for professional development, that extended time instructional activities (before-school, after-school, Saturday schools, summer school) be developed, and that teacher mentoring is ongoing. Finally, the district must provide all students in school improvement schools the option to transfer to another public school not identified as needing improvement. Transportation to other schools must be provided by the district and may require the use of local funds to finance it (although the NCLB legislation specifically notes that districts do not have to spend any local dollars to implement the requirements of the law). If a school in program improvement fails to achieve AYP by the end of the first year, the school must make supplementary services available to all struggling readers.

If, at the end of the second year of school improvement, the school still does not meet AYP for all subgroups, then "corrective actions" are mandated by NCLB. Possible corrective actions include (U.S. Department of Education, 2002):

- Replace the school staff relevant to the failure.
- Implement a new curriculum.
- Extend the school day or school year.
- Significantly decrease management authority in the school.
- Appoint outside experts to advise the school.
- Restructure the internal organization of the school.

If the school fails to meet AYP in the year following the corrective actions, the district must, by the beginning of the next school year, do one of the following:

- Reopen the school as a public charter.
- Replace all or most of the staff, including the principal.
- Enter into a contract with a private management company to operate the school or allow state takeover of the school.

If a school in program improvement makes AYP for two consecutive years, then it is not subject to school improvement mandates.

Public School Choice under NCLB

As soon as a school fails to achieve AYP goals, students in that school must be provided with the opportunity to leave that school and attend another school that met AYP goals. The lowest-achieving children from low-income families must be given priority in such cases. The district must pay for this transportation, which will likely leave less money to fund instructional interventions. Once a child has transferred, that child may remain in that school until the highest grade offered is completed. If, however, that child's original school achieves AYP goals and is no longer identified as needing school improvement, the district is no longer required to provide transportation to the choice school.

Although mandates on choice seem clear, there is a significant problem in some districts and even in whole states in fulfilling these mandates. In Florida, for instance, 87 percent of the schools in 2005 failed to make AYP and thus pupils in those schools had to be provided the opportunity to transfer to one of the 13 percent of the schools that met AYP. A similar situation existed in most large urban school districts. But when 9 out of 10 students attend schools needing improvement, it is impossible for sufficient space to exist at the few schools meeting the AYP goals to enroll all the students who might want to transfer.

Additionally, even if the space existed, school districts would almost necessarily have to move the teachers from the failing schools to the achieving schools just to have sufficient instructional staff. Is it likely that changing the building a teacher teaches in will automatically result in far more effective teaching by that teacher?

Supplementary Services under NCLB

Supplementary services are defined as "tutoring and other supplemental academic services" that are (U.S. Department of Education, 2002):

- in addition to the instruction provided during the school day
- of high quality, research-based, specifically designed to increase the achievement of eligible children

These supplemental services may be provided by the district, but the district must give parents a list and description of other qualified supplemental services providers, including for-profit providers (e.g., Sylvan Learning Centers, private tutors, etc.). The district must also provide transportation to and from supplemental services up to a targeted maximum amount of dollars (the formula for funding transportation is complicated and involves a percentage of the amount of federal dollars received and a percentage of the local funding).

The big shift here is that supplemental services are to be provided outside the school day in after-school, before-school, and Saturday school programs. This may ameliorate the problem of pulling children out of their classrooms during the day (which always results in loss of classroom instructional time), but it may exacerbate the problem of fragmentation of instructional support. In other words, if current programs seem fragmented, with little coordination between, for example, special education reading lessons and classroom reading lessons, then how difficult will it be to ensure a coherent instructional plan for children being served in supplemental services provided in an after-school program offered by a for-profit provider?

Other Problems with NCLB Mandates

One huge problem with the several mandated corrective actions is the almost complete lack of research support for any of the options. There exists no viable research indicating that replacing a principal, or replacing some instructional staff (let's say replacing the remedial reading or special education teachers), reliably enhances student achievement. Likewise, there is no research indicating that purchasing a new reading program reliably raises reading achievement (Berends et al., 2002; McGill-Franzen et al., 2006). Nor is there evidence that turning a school over to a management firm or to the state or that creating a charter school reliably improves achievement. No research indicates that school choice improves student reading achievement (Ravitch, 2010). It seems odd, given the emphasis on using "research-based" interventions in NCLB, that so many mandates are offered that have so little evidence of success in improving achievement.

The federal government, however, has been promoting such unsupported mandates for at least two decades now. Beginning with the Comprehensive School Reform and Development Act (CSRDA) of 1990, continuing with the Reading Excellence Act (REA) of 1998, and now with NCLB, the U.S. Department of Education, following legislative mandates of Congress, has been attempting to improve reading achievement by mandating or "incentivizing" the use of "proven" programs. The problem is that there are no "proven" programs (Allington & Nowak, 2004). At least there are no existent programs that reliably increase student achievement across multiple sites.

Berends and colleagues (2002) found that when schools adopted one of the "proven" program models touted by the federal government (e.g., Success for All, America's Choice, Roots and Wings, Comer Schools, Accelerated School, etc.), achievement did rise in some schools but there was no consistent pattern of improvement across the schools engaged in such adoptions. But this is no surprise for reading researchers familiar with more than 50 years of research on the effects of programs on student reading achievement.

In fact, in the largest national study comparing different reading programs, a study conducted 45 years ago, the authors (Bond & Dyskstra, 1967) concluded, "Future research might well center on teacher and learning situation characteristics rather than methods and materials" (p. 123). They reached this conclusion because their study indicated that all programs worked somewhere and none worked everywhere. Their arguments were similar to those made more recently by Berends and colleagues (2002), who concluded that what mattered was "local capacity"—the teachers and the workplace context in which the teachers worked. In both studies, and in others of smaller scale, programs, whether defined as commercial reading series or as systemic reform models, had small and variable impacts on student reading achievement. But in no study were any programs of any type identified that reliably raised reading achievement from site to site. None.

A second major flaw in NCLB is the heavy reliance on testing, especially standardized testing, as the basis for estimating school effectiveness and student achievement (Papay, in press). The use of standardized tests is less problematic for evaluating school improvement than for estimating individual achievement growth, assuming that the testing data are uncontaminated. But problems remain in both cases.

Group standardized achievement tests are simply not designed to provide estimates of individual reading growth. Almost any technical manual accompanying the most widely used tests clearly states this fact. Every major research and measurement organization, as well as the National Research Council, has opposed using standardized test data in making decisions about an individual student's achievement—for example, who to flunk or who to place in remedial reading classes. One major reason schools employ psychologists to assess children recommended for special education services is the recognized limitations of group achievement tests to provide reliable and accurate estimates of individual achievement. But NCLB mandates standardized testing to identify AYP attainment, to identify children for supplemental services, and so on. The achievement tests available today simply cannot provide the sort of information school districts need to make decisions about students.

Even though group achievement tests are more appropriate for estimates of school effectiveness (because the error inherent in the tests is largely ameliorated with large samples of student scores), this stance is supportable only when the achievement data are uncontaminated. Unfortunately, there are several sources of test contamination that impact the achievement data in just about every school in the nation.

Robert Linn (2000), in his presidential address at the American Educational Research Association noted, "Assessment systems that are useful monitors lose much dependability and credibility for that purpose when high stakes are attached to them" (p. 14). Because of test contamination from various sources, tests that

might provide useful and reliable estimates of school effectiveness are made largely useless for that purpose. There are three common sources of test contamination that undermine the usefulness of tests for measuring AYP and for determining which schools are in need of improvement and which are not: flunking, summer reading loss, and test preparation.

Flunking

As high-stakes testing expands, so does flunking (Allington & McGill-Franzen, 1992). Although research has shown flunking to be an ineffective and expensive response to academic difficulties (Shepard & Smith, 1989), this seems not to have deterred politicians from mandating that lower-achieving children be retained in grade. Several states and large urban districts, including New York City, now require the flunking of students who fail to attain an achievement standard. But flunking is not just bad for kids—it also contaminates the accountability system. In 2003, some 30,000 third-graders were flunked in Florida. Subsequently, the fourth-grade state test scores rose. Imagine that—just removing the lowest-scoring children allows a governor to proclaim an education reform is working.

But, of course, the schools that flunked all those students were not actually doing a better job of teaching children to read, even if their fourth-grade test scores did rise. The question is what would the scores have looked like if all those flunkees had had their scores included in the fourth-grade testing results? A year later, almost half of those who flunked the previous year were again scheduled to be retained in grade because their reading achievement had not improved. So much for the benefits of being retained.

Summer Reading Loss

Poor children lose ground in reading during the summer vacation period. Cooper and colleagues (1996) conducted a meta-analysis of studies of summer setback. They found that poor children begin school in the fall about three months behind where they were when they left school for summer break. In contrast, more advantaged students actually gain a bit during the summer months, starting school a bit ahead of where they were when summer vacation began (Allington & McGill-Franzen, 2003). The accumulating effect of this summer setback means that poor children during their elementary school years may fall as much as two years behind more advantaged students, even if the instruction during the school year produced identical reading growth in both groups. In two large-scale studies it was estimated that 80 percent of the rich/poor reading achievement gap was created during the summer months, not during the school year (Alexander et al., 2007; Hayes & Grether, 1983). In other words, both economically advantaged and disadvantaged children made the same reading progress during the school year but the poor children slid backwards every summer, thus creating most of the reading achievement gap.

However, if we provide poor children with books they want to read and books they can read during the summer months, we can ameliorate summer reading setback (Allington, McGill-Franzen, Camilli et al., 2010; Kim & White, 2008). In fact, providing a dozen or so free books to primary-grade children every summer improved reading achievement as much as attending summer school did! Providing five or six books to older elementary students achieved the same result. However, hardly any schools send poor kids home with a supply of books to read over the summer months. Thus, summer reading setback potentially contaminates reading test scores because it works to lower the reading gains children enrolled in high-poverty schools make.

This is where the accountability system is contaminated, at least if the system is intended to measure how effective a school's instruction might be. In other words, schools enrolling many poor children will find it much more difficult to achieve AYP than schools with few poor children. This is true even if the high-poverty schools are as effective as the middle-class schools. Annual AYP testing assumes that any growth, or loss, can be attributed to the school. But summer setback occurs when the school is not in session. High-poverty schools will necessarily have to address students' summer reading because the evidence suggests that when children practice reading during the summer months, their reading proficiency actually improves. When children don't read during the summer months, their reading skills decline.

Test Preparation

Test preparation, if it raises scores without actually improving reading achievement, is another factor that might contaminate accountability. However, there is little evidence that the sort of test preparation that seems to pervade many schools actually increases test scores. Guthrie (2002) notes that almost all the variance in test scores is accounted for by reading ability and general knowledge of the world. Test preparation might produce a small benefit if it works to ensure that students are familiar with the test format, but too much practice on formats produces careless errors.

The best guideline for test preparation would seem to be to practice a couple of days before the test to familiarize students with the test format and to introduce, or review, general test-taking strategies. But daily periods of test preparation across the school year seems more likely to result in lower performances because most test preparation involves little, if any, teaching of useful reading strategies or development of word knowledge. Until research exists demonstrating the positive effects of test preparation on reading achievement, schools should avoid such activities except at the most modest scale.

The Future of NCLB

As the NCLB sanctions continue to be imposed, it is likely that changes in the law will come. The U.S. Department of Education has already changed some provisions relating to pupils with disabilities and with the formula for calculating AYP. The current administration has proposed extending the testing requirements into the high school years. But no matter the changes, NCLB still represents a continuation of a long-standing but failed federal policy. Missing from the NCLB is the recognition that accountability comes only with autonomy (Allington, 2002b). When required to follow a script or implement a mandated program, neither teachers nor administrators are likely to assume much responsibility for any failure that ensues. Only when federal policy shifts to providing autonomy for teachers and principals—autonomy to elect the instructional plan—will we likely see the sorts of improvement policy makers hope for. I see some hope that federal policy makers may be moving in this direction in the recent creation of the Response to Intervention initiative.

Response to Intervention

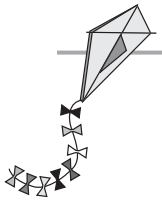
As part of the 2004 reauthorization of the Individuals with Disabilities Education Act (IDEA), Congress created a new response to intervention (RTI) initiative aimed at reducing the numbers of students identified as "pupils with disabilities." Because approximately 80 percent of all pupils with disabilities exhibited reading difficulties, the RTI initiative targeted accelerating reading development in the early grades for children who began school behind their peers. Roughly two-thirds of all kindergartners entered that grade already knowing the names of all the letters of the alphabet. Roughly one-third of entering kindergartners did not know this. Five years later, roughly two-thirds of fourth-grade students achieved the basic level of reading proficiency on the NAEP; one-third of the students fell below the basic level, or substantially behind their peers. It seems likely that those fourth-graders in the bottom one-third were the same one-third of students who entered kindergarten not knowing the alphabet.

School districts are now allowed to take the equivalent of 15 percent of their total special education expenditures and use that money to fund RTI. There is no role in the RTI regulations for special education personnel nor for school psychologists—at least not until after the intervention has been conducted and the children who failed to have their reading development accelerated are identified. At that point, then, these students become pupils with disabilities and are from that point onward served through special education resources.

However, as Johnston (in press) has noted, the RTI regulations include both a plan to provide added research-based reading instruction for struggling readers in an attempt to reduce the numbers of struggling readers and a new plan for identifying pupils with learning disabilities. In far too many school districts the emphasis has been on the latter rather than on planning powerful interventions that bring struggling readers up to grade level. One other factor for this problem might be that the majority of the books now available on RTI have been written by school psychologists with an emphasis on identifying the pupils with learning disabilities.

There are now several professional texts with an emphasis on providing high-quality reading interventions (Allington, 2009c; Howard, 2009; Johnston, 2010). Whether the RTI initiative will accomplish Congress's intended goal, reducing substantially the number of children identified as having a learning disability will depend largely on how well schools implement powerful intervention plans. But as Vellutino and Fletcher (2005) have noted, "Finally, there is now considerable evidence, from recent intervention studies, that reading difficulties in most beginning readers may not be directly caused by biologically based cognitive deficits intrinisic to the child, but may in fact be related to the opportunities provided for children learning to read" (p. 378). In other words, there simply do not seem to be any children who meet the current definition of learning disabled or dyslexic. That is because such children can have their reading development accelerated but only when they receive one-to-one expert tutoring or one-to-three very small group intensive expert reading instruction. When schools do not have kindergarten tutorial interventions, when they do not have first-grade tutorial interventions, when they do not have expert reading teachers providing these reading interventions, the result is a large number of children who become labeled as learning disabled or retained in grade or both. By and large this is not a result of having too little money to address these problems; it is more simply that most schools spend the money they have on lots of things that have never been supported by the research.

What Are Characteristics of Scientifically Based Reading Research?



When reviewing research findings to determine whether the research met the four criteria specified in federal legislation (listed in bold in the list that follows), readers may want to ask themselves questions about how well any particular study meets each of the criteria. Examples of the types of questions that could be asked about each criteria are included here.

- **Use of rigorous, systematic, and empirical methods.** Does the work have a solid theoretical or research foundation? Was it carefully designed to avoid biased findings and unwarranted claims of effectiveness? Does the research clearly delineate how the research was conducted, by whom it was conducted, and on whom it was conducted? Does it explain what procedures were followed to avoid spurious findings?
- **Adequacy of the data analyses to test the stated hypotheses and justify the general conclusions drawn.** Was the research designed to minimize alternative explanations for observed effects? Are the observed effects consistent with the overall conclusions and claims of effectiveness? Does the research present convincing documentation that the observed results were the result of the intervention? Does the research make clear what populations were studied (i.e., does it describe the participants' ages, as well as their demographic, cognitive, academic, and behavioral characteristics?), and does it describe to whom the findings can be generalized? Does the study provide a full description of the outcome measures?
- **Reliance on measurements or observational methods that provided valid data across evaluators and observers and across multiple measurements and observations.** Are the findings based on a single-investigator single-classroom study, or were similar findings observed by multiple investigators in numerous locations? What procedures were in place to minimize researcher biases? Do observed results hold up over time? Are the study interventions described in sufficient detail to allow for replication? Does the research explain how instructional fidelity was ensured and assessed?
- **Acceptance by a peer-reviewed journal or approved by a panel of independent experts through a comparably rigorous, objective, and scientific review.** Has the research been carefully reviewed by unbiased individuals who were not part of the research study? Have the findings been subjected to external scrutiny and verification? Has the study been published in a peer-reviewed research journal?

Finally, as mentioned earlier, you can now go to the What Works Clearinghouse website to find out what research says about reading programs. Each WWC entry lists the studies they reviewed along with an indication of why certain studies were rejected from consideration by the WWC reviewers. Basically, these studies all suffered at least one major deficiency that made the findings less than unbiased and less than reliable. The truth of the matter is that I, or almost anyone else, can design a "study" that shows that your favorite reading program "works." In fact, many such studies have been done, but most often those studies violated some fundamental

aspect of good research design. Thus, no matter how many studies have been done, if they are poor studies they don't deserve recognition. To see a number of such studies just visit the websites of the publishers of almost any reading program. But always remember: Buyer beware.

Thinking about Research in Reading

The first guideline focuses on the *use of rigorous, systematic, and empirical methods* in the design of the study, which is not particularly surprising. After all, research has traditionally been an empirical adventure. But several potential problems are created here. Take the issue of *bias,* for instance. If a program developer closely monitors the implementation and the evaluation, is there an unintended biasing effect? In other words, would program developers pay more attention to implementation detail than an independent evaluator or the professional staff of a school district who also decided to implement the program and gauge its effects? My guess is that, yes, probably the developer would pay more attention. If so, are the effects the developer achieves reliable? That is, can they be achieved by others? Realize that the added attention the developer pays to implementation may not come from any ego- or profit-driven motive but, rather, from a clearer understanding of just how the program is supposed to work. If the program involves providing specific training to the teachers involved, can any two staff development providers actually offer identical training? And what if the teachers at your site are less experienced and have larger classes than the teachers at the developer's site? The point here is that "rigorous and systematic" methods often have to be adapted from site to site. The old saying that "a teacher is a teacher is a teacher" just isn't true any more than suggesting that classrooms and schools are all largely comparable.

The difficulty in designing rigorous, systematic research studies in real schools is typified in the armful of "adequacy" reviews that have been published (e.g., Coles, 2003; Lysynchuk et al., 1989; Pressley & Allington, 1999; Swanson et al., 2003; Troia, 1999). These reviews share a single common feature: All note how few published studies meet rigorous and systematic criteria. Swanson and Hoskyn (1998), for instance, reviewed over 900 studies of instructional interventions with children identified as learning disabled. Of these studies, only 180 met minimal criteria for rigor, and *fewer than 10 percent were rated as exhibiting high-quality research methodology.* Troia (1999) reviewed 39 studies of phonemic awareness interventions and noted that fewer than a quarter met even two-thirds of the criteria of rigorously designed research. The National Reading Panel (2000) located only 38 studies of phonics instruction that met their criteria for rigor. This same finding is why the What Works Clearinghouse has been dubbed the "Nothing Works Clearinghouse" by some educators. The truth of the matter, however, is that very few studies, if any, support the majority of decisions that principals and teachers have to make every day.

Much messiness in educational research stems from the problem of achieving purely random assignment of subjects. It would be difficult, if not impossible, to approach a school system and ask for participation in a research study that required all teachers and children to be randomly assigned to buildings across the district. But without such random assignment it is impossible to control for school, teacher, and community effects that might bias the outcome. In virtually all funded research, both teachers and children (actually their parents) must volunteer to participate. What about the teachers who don't volunteer? Are they comparable to the volunteers? In other words, does not including these teachers bias the results? Would teachers who were better teachers be more likely to volunteer? That would bias the effect of the intervention in a positive direction. Would parents of higher-achieving students be more likely to return permission slips than parents of lower-achieving students? That would also create a positive bias. Do some schools have a larger supply of better teachers? Or a larger number of low-achieving children?

In order to conduct a "true experiment," researchers must attempt to eliminate such bias in their subject sample. *Random selection* is the historical strategy for eliminating such bias. In true random assignment, every teacher or student or classroom is randomly selected from the population either to participate in the intervention or to serve as a control participant. The control participants do not receive the special treatment. But as anyone who works in schools knows all too well, getting teachers and parents to agree to random assignment to schools and classrooms is simply not feasible. Even getting schools to randomly assign students within a building has not been easy.

Think of the issue another way. If you wanted to achieve the best effects from an intervention, wouldn't selecting the interested volunteer teachers as the intervention teachers be desirable? That is just what most schools (and many researchers) do when they field test an intervention. But in a rigorous research study, such biased selection would violate this first principle. At the same time, following federal human subject protection guidelines, required for virtually all federally funded studies, means that participating teachers provide informed consent. In other words, teachers must be informed about the study and have the opportunity to decline to participate. We should not be surprised that teachers who see the proposed intervention in a more favorable light are more likely to volunteer. Thus, it seems that much research presents a best-case scenario—which is how the intervention works when teachers volunteer to try it.

Participating in an unbiased research study can create public relations problems for a school district. Consider, for instance, the sorts of parental concerns that could arise even with no random assignment of teachers. If some children are randomly selected to participate in an early reading intervention, say, and other children with similar needs are selected to be the unserved control students, parents of this latter group will undoubtedly object to this lack of services. Or if some

randomly selected children receive a tutorial intervention and other similar children are assigned to work in a small group with a paraprofessional, parents can object and with reason. Also imagine how much more difficult the situation becomes when some classrooms are offering the special program while others are not. Rigorous, unbiased scientific research is an ivory tower standard that is simply very hard to accomplish in the real world of schools, teachers, and children.

The second guideline asks whether *the data analyses were adequate to test the stated hypotheses and justify the general conclusions drawn.* In an ideal world, every study would have a randomly selected group of teachers and children who received the experimental intervention and another randomly selected group of teachers and children who did not. Everything about the instruction offered would be identical, except one group would participate in the intervention and the other would not. The question is, When do we fit the intervention in if instruction is to be otherwise comparable? Would some children simply stay in school longer each day? That doesn't work because then we couldn't decide whether it was the intervention instructional design or just adding more instruction, regardless of the type, that led to any observed achievement effects.

However, participating in a special project has been observed to raise scores even if no real intervention is offered. This has been dubbed *the Hawthorne Effect.* This effect was first noticed in a manufacturing plant in Hawthorne, New Jersey, many years ago. Workers there were more productive when they were told they were being studied as part of a special project even though no actual experiment was conducted. To guard against the "added instruction" and Hawthorne effects, rigorous research design typically attempts to provide some other special intervention to the control group but an intervention thought not to have any real impact on, say, reading achievement.

So, a researcher might add a reading tutorial to the daily schedule of a group of randomly selected lower-achieving readers to assess its effects. At the same time, another group of randomly selected lower-achieving readers would receive a handwriting tutorial. In this case, the Hawthorne Effect is effectively nullified. Both groups of children receive a tutorial. But the added instruction problem still exists. One group received additional reading instruction, the other didn't. So why would anyone be surprised if the students receiving the added reading instruction had higher reading achievement at the end of the study?

To counter the added instruction problem, the researcher might offer two types of tutorials, both targeted at reading improvement. In one case the children might be tutored with an emphasis on developing decoding strategies—an Alphabetic Phonics tutorial. The other group might receive a different focus, perhaps reading fluency training with an emphasis on rereading texts until a certain fluency level has been achieved. In such cases, both the Hawthorne Effect and the added instruction problems are effectively countered, assuming that both groups of students

and their tutors were randomly selected. If the children participating in one of the tutoring interventions record higher reading achievement at the end of the year (and perhaps for years to come), then with such a design it would be possible to consider that the observed effects were neither biased nor chance effects but real achievement gains attributable to differences in the effectiveness of the interventions.

But what happens if the effectiveness of the intervention is measured on a test of the ability to pronounce nonsense syllables? Does the selection of that outcome measure bias the outcome? Would significantly higher scores on a nonsense word test suggest that the decoding intervention was more effective—at fostering improved reading achievement? What if the outcome measure was a test of reading fluency? Is a test of fluency a more appropriate test of reading achievement than a test of nonsense word pronunciation? Would a test of fluency be biased toward the fluency intervention? How about a test of spelling? Or retelling of a narrative read silently? What if the decoding intervention improved nonsense word pronunciation but had no effect on fluency, and the fluency treatment produced the opposite result? Would such a finding be unexpected? Gamse and colleagues (2009) found that children in Reading First schools did do better at reading nonsense words than children in the control (non-Reading First) schools but that gain did not lead to improved reading achievement in Reading First schools. Again, given that teachers in Reading First schools spent more time on nonsense word decoding than the other teachers did, it is not surprising to me that the Reading First kids did better at reading nonsense words. But it also isn't surprising to me that better nonsense word reading did not lead to improved reading achievement. As far as I know, it never has! All of which begs the question: Is your school still teaching and testing nonsense word reading? And if so, why?

Thus, how the effects of an intervention are evaluated makes a huge difference in conclusions about effectiveness. This is an important issue because one criticism was that too much of the intervention research focused on developing decoding skills. Although there are reported effects on nonsense word pronunciation, these studies less often reported positive effects on other assessments of reading achievement (Allington & Woodside-Jiron, 1999; Cunningham et al., 1999; Gamse et al., 2009; Pressley & Allington, 1999). In other words, many phonics interventions demonstrate improved pronunciation of nonsense words but no improvement in reading achievement (e.g., reading fluency and comprehension).

There are also concerns about what sort of reading assessment was used—experimenter-developed, nonstandardized commercial batteries, or standardized commercial assessments? Swanson and Hoskyn (1998) reported that the studies they reviewed that used standardized assessments of reading routinely produced smaller gains than the studies that reported results on experimenter-developed tests and tests of subskills. In other words, it is easier to design a study that produces achievement effects, especially on experimenter-developed assessments, than a

study that produces effects on standardized assessments. But that shouldn't be surprising. However, it should be a concern when considering the effects of any intervention because what we are attempting to do is to create better readers.

Another issue to consider here is the size of the effect observed. Historically, tests of statistical significance have been used to estimate the reliability of differences in achievement between groups. But a test of statistical significance only tells us that observed differences did not occur by chance, regardless of how small that difference is. So, with a large enough sample of students, even small differences in achievement can turn out to be statistically significant. But when do such differences become educationally significant?

Venezky (1998) reported on a reanalysis of the reading achievement in schools that had adopted the Success for All (SFA) program. He noted that fifth-grade students in SFA schools had average reading achievement grade levels of 3.6, whereas the students in the control schools had a 3.2 average reading level. The four-month difference in reading achievement was statistically significant, but Venezky asked whether—after six years of an intervention—such a difference was educationally significant or cost-effective. He suggested that it seemed less important that intervention results be compared statistically to control groups and more important that interventions also be evaluated against a fixed standard: How many students achieved the state standards, achieved grade-level performance, and became avid, voluntary readers?

Finally, it is important to consider the generalizability of the results. Does the population in the study represent the diversity of teachers and students found in U.S. schools? This is, of course, a standard that would be almost impossible to achieve in a single research project. Nevertheless, it is a useful question to consider. If the study was conducted in New York state, where all teachers must earn a master's degree within five years and your school is located in a state with less rigorous standards for teachers (and perhaps many teachers working with emergency credentials), can the results be generalized from one location to another? If the teachers who implemented the intervention were volunteers, can we expect the same effects from teachers who were mandated to implement an intervention? If the study was conducted in a school located in an upper-middle-class suburb, can the results be generalized to schools in any neighborhood? What if the study school had a 20:1 student–teacher ratio and your school has a 28:1 ratio? Generalizability rests, in large part, on comparability of populations—both students and teachers. Without rich information on the community, the teachers, the school context, and the students, it is difficult to judge comparability.

The third guideline asks whether the researcher *relied on measurements or observational methods that provided valid data across evaluators and observers and across multiple measurements and observations.* This could be considered a "fidelity of implementation" guideline. Did the researcher provide evidence that the intervention was

actually implemented? Did observers monitor implementation? Was the effectiveness of the intervention related to quality of the implementation? Consider, for instance, California's experience: Almost simultaneously, schools in California were provided funds to (1) reduce class size in the primary grades, (2) provide primary grade teachers with staff development on teaching phonics, and (3) provide new phonics-emphasis curriculum materials. How might a researcher sort out the separate effects of these three different, but simultaneously enacted reforms?

There actually is almost no way one could attribute improved achievement to any of the three reforms without some careful observation of the staff development provided and the instruction offered before and after participation in the staff development. On the other hand, we now know that all those changes in California did not result in improved reading achievement. But it is still unclear why that is so.

In other words, to attribute achievement changes to participating in the staff development intervention, you would have to be able to document how classroom instruction changed after the staff development and then link those changes to improved student achievement. If some teachers changed their instruction substantially and others did not, then you might expect some classes to have substantial gains and others to make modest improvements, if any. In such a case, claims linking improved achievement to participation in the staff development might be reasonably made. Now, if the changes in instruction involved a greater frequency of the use of the new phonics materials, then perhaps claims could be made about their influence. But there is still the problem of sorting out the effects of the class-size reduction reform. Smaller classes produce better achievement even without professional development or new curriculum materials of whatever ilk (Achilles, 1999).

Note that in this example there is no control group per se. All teachers participated in the staff development. Judgments about the link between staff development and improved achievement are supported by linking differential implementation to differential patterns of achievement growth. Finding any such studies, however, is enormously difficult. This may be because of the expense involved in having observers in classrooms both before and after participation. It may be because policy makers rarely produce policies that are easily studied. Or it may be because policy makers seem rarely interested in evidence of the effects of their policy making (Allington, 2001). So, rather than provide funding for research, politicians just assert that their favorite policy was the basis for any improvements observed (with no evidence, scientific or otherwise, to support the assertion).

So far we have not considered the problem of ensuring that different observers actually observe and record the same teaching behaviors in the same ways. This is called *interrater reliability.* In other words, two raters watching the same lesson would rate it nearly identically on some observational scale. Similarly, we haven't discussed constructing an observational system that has been shown to focus on the key

features of instruction—the features that produce the higher achievement. More often than not, research studies that include a monitoring of the fidelity of implementation fall short on these two criteria: demonstrations of interrater reliability and demonstrations of the predictive validity of the observational scale.

Now all of this seems overly technical and more than a bit nitpicky. But consider the claims of effectiveness that have been made for any number of methods, materials, and programs and then consider the evidence available to support those claims. Truth be told, it is difficult to find evidence that meets the What Works Clearinghouse criteria for any educational innovation. That shouldn't be surprising since even the highly lauded medical research—often held up as the ideal for education—has suffered through charges that much of the research available used only men as subjects; used populations that underrepresented minorities, people living in poverty, and people living in rural regions; excluded the obviously healthy in their samples; or was potentially biased because the funding for the research came from the drug company marketing the treatment. The recent flurry of contradictory medical research on the role of salt in one's diet, the utility of mammograms, the benefits of red wine, and so on, points to the difficulty that even medical science has in delivering "rigorous, reliable scientific research" that achieves consensus findings (Allington, 2004c).

Far too much educational intervention research falls short on the fidelity of implementation criteria. That is, few studies even attempt to estimate whether the intervention was effectively implemented. The cost of such a demonstration and the lack of demand for such evidence combine to create an environment whereby it is asserted that an intervention was implemented and then assertions about the effects, or the lack of them, are also made. In order to understand the effects of an intervention it is necessary to gather information on its implementation. This seems especially true in education where any number of studies have shown that rather few features of any intervention ever studied were actually implemented as imagined by the developer and few of the implemented changes survived over the longer term (Berends et al., 2002; Datnow & Castellano, 2000; McGill-Franzen, 2000).

The final guideline suggests that when papers reporting research on method, materials, and programs have gained *acceptance through publication in a peer-reviewed journal or approved by a panel of independent experts through a comparably rigorous, objective, and scientific review,* more confidence can be placed in the findings. I agree. But this is hardly a fail-proof test. Consider that the majority of the studies reviewed by Lysynchuk and colleagues (1989), Pressley and Allington (1999), Swanson and Hoskyn (1998), Troia (1999), and the National Reading Panel were published in peer-reviewed journals. Yet, typically fewer than half of the published studies reviewed met the minimum quality standards. Suffice it to say that there are many published, but less than rigorously designed, studies in educational and psychological research journals.

Part of the reason for this has been alluded to previously. It is tremendously difficult and, typically, expensive, to design and carry out well-designed educational intervention research. Given how financial support for educational research has eroded over the past 30 years, few school districts or state education agencies or publishers actually fund any research on methods, materials, or programs. That said, no one should be surprised that far too many educational research projects cut corners (and costs) in ways that impact the quality of the results—at least in terms of the confidence we can place in the reliability and generalizability of the results.

In addition, every year more educational and psychological journals and magazines appear in the marketplace. These magazines need articles to fill their pages. Thus, the past 30 years have seen a veritable explosion of lower-quality research. Relatively few journals can be considered high-quality research publications where the peers doing the review are established and recognized educational researchers. Too many educational publications exhibit precisely the opposite attitude of that exhibited by the popular media. While the popular media seems to focus almost exclusively on "bad news" stories about education, educational magazines and journals focus primarily on "good news" stories. Think about the educational magazines and journals you read. How many articles in these publications reported on the failure of a reform or an intervention?

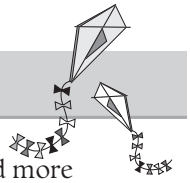
High-Quality Educational Research Journals

There are literally a hundred or more journals and magazines that publish educational research. However, there are but a handful of journals that require rigorous peer review. While no list can be comprehensive, below are my nominees for the journals most likely to publish high-quality studies or reviews of research:

Review of Educational Research
American Educational Research Journal
Journal of Educational Research
Journal of Educational Psychology
Journal of Literacy Research
Reading Research Quarterly
Elementary School Journal
Educational Evaluation and Policy Analysis

Two plain-language books that discuss U.S. educational achievement patterns in detail are Gerald Bracey's *Setting the Record Straight* (Heinemann) and Richard Rothstein's *The Way We Were?* (The Century Foundation). Both books cover broad academic achievement patterns for K–12 and include data on various other schooling issues such as flunking, teacher accountability, minority achievement, and so on.

Summary



It is important, I believe, that educators become better informed and more critical of claims of educational effects—positive or negative. As a profession we need to become more skillful at reading the promotional claims and the research assertions for educational interventions. I think we need to become more informed consumers of methods, materials, and programs. Claims of effectiveness have increased geometrically now that "research-based" instruction sits in the spotlight. But every claim needs to be examined with a skeptic's eye while applying the general guidelines offered here.

Let me offer one example. Currently, many claims of effectiveness are being made for reading programs that include "decodable" texts (e.g., "Nan can fan the man"). The proponents suggest that it is a phonics emphasis with the accompanying use of "decodable" texts that make such programs effective. But, until recently, there has not been not a single study that systematically manipulated the use of decodable texts—texts where almost all the words are pronounceable given the letter–sound associations that have been taught—including studies examining the effectiveness of phonics programs (Allington & Woodside-Jiron, 1998). Claims about the utility of decodable texts are not supported by the research now available (Jenkins et al., 2004). But this lack of research has not inhibited proponents of a more code-emphasis—or phonics-emphasis—curriculum. Indeed, several states have now mandated the use of decodable texts and in each case assert, incorrectly, that their policies are "research-based."

Likewise, the widely distributed booklet *Put Reading First* (Armbruster et al., 2001), along with the "scientific" entrepreneurial guidelines widely used to select reading programs for use in Reading First schools (Simmons & Kame'enui, 2002), both include decodable texts as one of several non–research-based criteria for identifying "scientific" curriculum materials. The fact that your federal government has promoted both of these guides is a cause for concern. If the federal education agencies cannot reliably report what the research actually says, given enormous resources, how can teachers be expected to accomplish this feat?

All of this may lead you to think that educational research is not going to be very helpful in designing higher-quality reading instruction.

But you would be wrong. We have learned an enormous amount about the characteristics of more effective reading instruction (Allington, 2009b; Pressley, 2006; Taylor et al., 2003, 2005). To use these findings, however, you have to move beyond the current fixation on methods, materials, and programs. When we ask about which method, material, or program is most effective, we ask a question that, literally, cannot be answered by referring to the research. As noted throughout this chapter, virtually every method, material, and program has accumulated some evidence that "it works!" But the evidence is often contradicted by other evidence.

In designing more effective reading instruction, we will need to look to the research for larger issues than answers to questions about particular methods, materials, and programs. There seems a simple, but often overlooked reason for this. The search for any "one best way" to teach children is doomed to fail because it is a search for the impossible (Cunningham & Allington, 2011).

A simple principle—children differ—explains why there can be no one best method, material, or program. This simple principle has been reaffirmed so repeatedly in educational research that one would think most folks would have noticed it by now. In addition, anyone who grew up with siblings or who has more than one child of her or his own, knows from powerful experience that no two children are alike. Not even those from the same family gene pool. What, then, can you say about a classroom with 24 children from 48 sets of pooled genes?

A corollary principle—teachers differ—has been largely ignored as well, even though, again, we have lots of research evidence on the issue. In other words, no teachers are exactly the same. We've learned just how hard it is to get teachers to teach "against the grain"—to teach in ways that contradict their beliefs and understandings about teaching, learning, and reading and writing. If you want an intervention to fail, mandate its use with a school full of teachers who hate it, don't agree with it, and are not skilled (or planning to become skilled) in using it. This is what Linda Darling-Hammond (1990) has called "the power of the bottom over the top" in educational reform.

Additionally, if we are ever to create schools where all children are developing reading proficiency normally, we will need schools where every teacher believes that is possible and is committed to providing the types of individual instruction that some children need. I worry when I read a paper such as the one I recently read. That paper (Scharlach, 2008) reported that two-thirds of the teachers studied did not believe they could teach all children to read. These teachers also provided instruction that was different from the one-third of the teachers who reported they could teach everyone to read. Basically, the teachers who felt they could not teach everyone to read set lower expectations and provided less and lower-quality reading lessons when compared to the teachers who believed they could teach all kids. Interestingly, teachers who felt they couldn't teach everyone to read cited various factors to explain their response. Those factors included low motivation of some children for reading, poor parental involvement, and the presence of learning disabilities, among other factors, for why they were not successful. The teachers who felt they could

The 100/100 Goal

Imagine that we could design schools where 100 percent of the students were involved in instruction appropriate to their needs 100 percent of the day. Imagine how different the achievement patterns of struggling readers might be. I will suggest that the 100/100 goal is, perhaps, the real solution for developing schools that better serve struggling readers.

Source: National Center for Education Statistics. *National assessment of educational progress.* Retrieved from http://nces.ed.gov.

teach everyone to read offered none of these factors as excuses. So, school faculties differ not just on how many expert teachers of reading are available but also on how many teachers believe they can be successful with every child.

In any school, then, you have a horde of students who differ in innumerable ways and a cluster of teachers who also differ in a myriad of ways. Expecting any single method, material, or program to work equally well with every kid in every classroom is nonsensical.

In the remainder of this book, I will address some of the lunacy of the current reading reform movement, especially the push to standardize reading instruction. Because federal legislation has set such a visible standard for using research to redesign reading instruction, I have attempted to develop a research-based argument for how we might best use what we have learned (from research) in the redesign of reading instruction in U.S. schools.

This book focuses on the converging evidence that is available on the features of reading instruction that really matter. I leave to others to debate particular methods, materials, and programs. In this book I develop a research-based framework for rethinking reading instruction generally, and particularly the reading instruction that we offer kids who struggle while learning to read.